PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-163140
(43)Date of publication of application : 07.06.2002

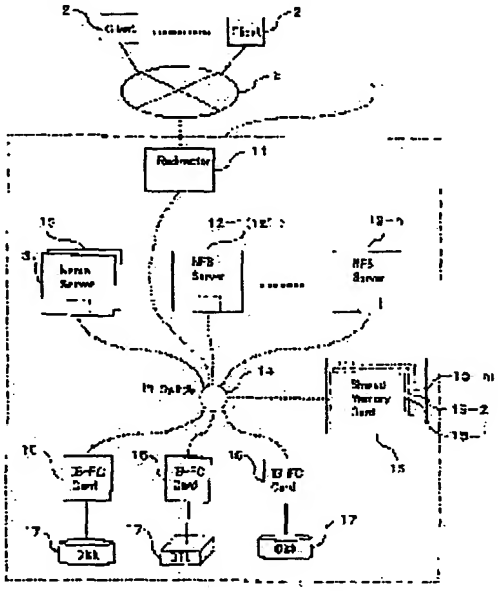
(51)Int.Cl. G06F 12/00
G06F 13/10

(21)Application number : 2000-359810
(22)Date of filing : 27.11.2000

(71)Applicant : FUJITSU LTD
(72)Inventor : OE KAZUICHI
NISHIKAWA KATSUHIKO

(54) STORAGE SYSTEM

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a storage system having a scalability capable of fully coping with the band expansion of a network at a low cost.
SOLUTION: This storage system is provided with a storage device 17 capable of storing file data, a plurality of file servers 12-1 through 12-n performing file processes in response to requests on file data to the storage device 17, a file server management node 11 managing the transfer processes of the file requests received from clients 2 via an external network 3 to the file servers 12-i (i=1 through n) and the response processes to the clients 2 for the file requests, and the internal network 14 communicatably connecting the storage device 17, the file servers 12-i, and the file server management node 11 together.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

* NOTICES *

CLAIMS

[Claim(s)]
[Claim 1] The storage which can memorize file data, and two or more file servers which perform file processing according to the request about this file data to this storage, The transfer processing to this file server of the request received from a client through an external network, The file server management node which carries out unitary management of the response processing to this client to this request, The storage system characterized by having offered the internal network which interconnects this storage, this file server, and this file server management node possible [communication], and being constituted.
[Claim 2] The storage system according to claim 1 characterized by connecting the name server which carries out unitary management of the file data name which this file server treats to this internal network.
[Claim 3] The storage system according to claim 1 characterized by connecting the shared memory with accessible this file server management node and this file server to this internal network.
[Claim 4] The storage system according to claim 2 characterized by connecting the shared memory with accessible this file server management node, this file server, and this name server to this internal network.
[Claim 5] A storage system given in any 1 term of claims 1-4 characterized by having offered the request analysis section in which this file server management node analyzes the content of this request, and the request transfer section which transmits this request to a specific file server according to the analysis result of this request analysis section.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[The technical field to which invention belongs] this invention relates to the storage system which enables sharing of file data by two or more clients by connecting with a desired network about a storage system.
[0002]
[Description of the Prior Art] As conventional technique of realizing sharing (only henceforth "file sharing") of the file data between two or more nodes (client) on a network For example, as typically shown in drawing 16 , a file server 200 is built on the desired networks 100, such as LAN (Local Area Network), using a Network File System (NFS:Network File System). A secondary storage 400 is connected to this file server 200 through the interfaces 300, such as SCSI (Small Computer System Interface; -- generally called "SCSI"). The method of realizing file sharing between two or more clients 500 in this secondary storage 400 is learned well.
[0003] However, the following technical problems occur by this method.
** Special skill is required to build and maintain a file server (maintenance).
** Extension (capacity, access performance) of a file server is not easy. Even if extensible, maintenance cost will increase by a file server being divided into plurality etc.
[0004] ** Special skill is required for the system construction and maintenance (maintenance) offered at the time of failure, and the costs for it also start.
NAS (Network Attached Storage) is proposed as a method of solving these technical problems in recent years. It is the system which the portion (refer to the dashed line frame in drawing 16) which consists of an above-mentioned file server 200 and an above-mentioned secondary storage 400 is equivalent to what was beforehand built as one storage system, connects with a network 100, and this NAS only performs an easy setup, and can realize file sharing, and special skill is unnecessary to construction and maintenance of a system (maintenance).
[0005]
[Problem(s) to be Solved by the Invention] However, also in such NAS, the technical problem that the scalability which can fully respond to band expansion (it will be about 10Gbps after (1Gbps and several years) in the present condition) of LAN which is progressing quickly now by the low cost is not obtained remains. That is, when it is going to correspond to band expansion of the network of a connection place, also in NAS, simply, an internal file server and an internal secondary storage will be extended, consequently a file server will be divided into plurality, and the secondary storage which each manages will also be divided.
[0006] That is, an above-mentioned file server 200 and an above-mentioned secondary storage 400 will be arranged in parallel and (independent) prepared. For this reason, it will be necessary to maintain a file server separately (maintenance), and maintenance cost will increase after all. It was originated in view of such a technical problem, and this invention aims at offering a storage system with the scalability which can fully respond by the low cost to band expansion of a network.
[0007]
[Means for Solving the Problem] In order to attain the above-mentioned purpose, the storage system (claim 1) of this invention The storage which can memorize file data, and two or more file servers which perform file processing according to the request to this storage, The transfer processing to the above-mentioned file server of the request received from a client through an external network, It is characterized by having offered the internal network which interconnects the file server management node which carries out unitary management of the response processing to the client to the request, and storage, an above-mentioned file server and an above-mentioned file server management node possible [communication], and being constituted.
[0008] Here, the name server which carries out unitary management of the file data name which the above-mentioned file server treats may be connected to the above-mentioned internal network, and the shared memory with accessible above-mentioned file server management node and above-mentioned file server may be connected to it (claim 3). (claim 2) In addition, when the above-mentioned name server exists, in addition to an above-mentioned file server management node and an above-mentioned file server, this name server can also be accessed at the above-mentioned shared memory (claim 4).
[0009] Moreover, it is desirable for the above-mentioned file server management node to offer the request analysis section which analyzes the content of the above-mentioned request, and the request transfer section which transmits this request to a specific file server according to the analysis result of this request analysis section, and to be constituted (claim 5).
[0010]
[Embodiments of the Invention] Hereafter, the gestalt of operation of this invention is explained with reference to a drawing.
(A) The explanatory drawing 1 of 1 operation gestalt is a block diagram showing the storage structure of a system (storage architecture) as 1 operation gestalt of this invention. The storage system 1 (only henceforth "a system 1") shown in this drawing 1 It is for realizing file sharing among two or more clients 2 connected to the external network (for example, Gigabit Ethernet (registered trademark)) 3. A redirector (Redirector) 11, two or more NFS servers (file server) 12-1 - 12-n, a name server 13, a shared memory (Shared Memory) 15, the IB-FC card 16, and a secondary storage 17 [a disk unit, a tape unit (DTL), etc.] It has offered. It has each of these components 11, 12-i, and the composition that 13, 15, 16, and 17 were mutually connected through the high-speed (interior) network [in FINI band (Infiniband)] switch 14 which is a 4-10Gbps (gigabit/second) grade.
[0011] Thus, if NFS server 12-i and a secondary storage 17 are connected to the internal network 14 if needed by taking the gestalt (clustering) which connects each components inside a system (a redirector 11, NFS server 12-i, a name server 13, secondary storage 17, etc.) in the internal network 14, it can extend easily, and the expandability (capacity, access performance, etc.) according to the transmission speed of the external network 3 etc. improves sharply.
[0012] Here, the above-mentioned redirector (file server management node) 11 is for carrying out unitary (concentration) management of the transfer processing to NFS server 12-i (i=1-n) of the various request messages (only henceforth a "request") received from the arbitrary clients 2 through the external network 3, and the response processing to the client 2 of the request origin to the request. That is, there is no need of maintaining them separately like before by existence of this redirector 11 though NFS server 12-i and a secondary storage 17 are extended like ****.

[0013] In addition, the above-mentioned "request" means the demand about the file data (only henceforth a "file") memorized by the secondary storage 17, for example, has the file manipulation demands (writing / updating / read-out) (file access request) to the substance of file data, meta-information access of other refer to the file name, etc. Moreover, from each client 2, only IP (Internet Protocol) address given to this redirector 11 can be referred to fundamentally. That is, from each client 2, this system 1 appears as one file server.
[0014] That a function which was mentioned above should be realized and to this redirector 11 For example, Gigabit Ethernet card 11a which equips the interface for external network 3 as shown in drawing 2 , In FINI band card 11b which equips the interface for the interior of a system (internal network), Network processor 11c for carrying out centralized control of the operation of

selves (redirector 11) including each of these cards 11a and 11b, Memory (primary storage) 11d for memorizing required software (program) and various data, when this network processor 11c operates etc. is mounted.
[0015] In addition, network processor 11c is connected with these components 11a-11d possible [two way communication] through PCI (Peripheral Component Interconnect) bus 11e. The above-mentioned network processor 11c here Transmission and reception (protocol conversion etc. is included) of the reply (response) message to the request sent and received between the internal network 14 and the external network 3, or its request etc., The analysis of the request (protocol) received from a client 2, the determination of the access file name based on the analysis result, the decision of destination NFS server 12-i of a receive request, etc. can be made, and the following control is also attained with this operation gestalt.
[0016] Namely, the request from a client 2 is analyzed, and it can control, or it can be controlled [the request about the same file can be assigned to the same NFS server 12-i, and] now so that file access competition does not occur between the NFS servers 12-i so that a request is uniformly transmitted to each file server 12-i (to for example, NFS server 12-i with a light load).
[0017] For this reason, if its attention is paid to the function of the important section, the following function parts are mounted in this network processor 11c.
(1) As the request analysis section 111 which analyzes the content of the request from a client 2 ****** (2) Transfer history of the request of the past by the function (3) request transfer section 112 as the request transfer section 112 which transmits a receive request to specific NFS server 12-i according to the analysis result of this request analysis section 111 (for example) By performing in the background the function (4) NFS server load surveillance demon (daemon) 115 as the transfer history Records Department 113 which records at memory 11d the function as the load Monitoring Department 114 which supervises periodically the loaded condition of each NFS server 12-i -- by this, the above-mentioned request transfer section 112 Based on the transfer history by the above-mentioned transfer history Records Department 113, transmit the request to the file of the same file name to same NFS server 12-i, or It becomes possible to transmit a receive request to low NFS12-i of a load based on the load surveillance result by the NFS server load surveillance demon 115 (load Monitoring Department 114).
[0018] Therefore, since a request can be uniformly transmitted to each NFS server 12-i while processing speed improves sharply, what a load concentrates on a part of NFS server 12-i, and causes an obstacle can be prevented certainly, and reliability improves it sharply. In addition, this network processor 11c carries out the cache of the reply message for example, to meta-information access to cache memory 11f (memory 11d is sufficient) of the interior (refer to drawing 3). When the request about meta-information access is received from a client 2, it is cache memory 11f () first. or if memory 11d is checked and hit, a reply message is created, and it will return to a client 2 side as it is (refer to drawing 4) -- (without transmitting to NFS server 12-i or a name server 13) it can also be made like
[0019] This structure is applicable not only about meta-information but file data. However, it is better to select file data with high access frequency by network processor 11c, and to be made to carry out the cache only of the file data to cache memory 11f (or memory 11d), since required memory space increased when it was made to carry out the cache of all the file data in this case.
[0020] That is, the above-mentioned cache memory 11f () Or the function as the cache section which carries out the cache of the reply message to the client 2 about specific file data with high request frequency is also achieved memory 11d. Network processor 11c in this case is it cache memory 11f () that it is a thing about the file with the same new request. Or it will also have the function as the response section 116 (refer to drawing 2) which returns the reply message which carried out the cache to memory 11d to a client 2.
[0021] Thus, if a response is returned by the redirector 11 about the information (meta-information, file data) that access frequency is high, without carrying out a cache by the redirector 11 side, and transmitting a request to NFS server 12-i, the speed of response to the client 2 to the information that access frequency is high will improve sharply, and the processing speed and the throughput as this system 1 will improve by leaps and bounds.

[0022] Next, the above-mentioned NFS server 12-i can access file processing (writing / updating / read-out) according to the request transmitted from the redirector 11 at internal network (internal network switch) 14 course at a secondary storage 17, can be carried out, or can generate the reply message for sending the file-processing result as a response to the client 2 of a requesting agency, and can be transmitted to a redirector 11, respectively.
[0023] Each NFS server 12-i in addition, in hardware For example, as shown in drawing 5 Offer interface card (IB-IF) 12c which equips interfaces (protocol conversion etc.) with CPU(Central Processing Unit) 12a, memory (primary storage) 12b, and the internal network 14, and it is constituted. By CPU12a's reading the NFS server software (program) memorized by memory 12b, and operating, the function as NFS server 12-i mentioned above is realized.
[0024] Now, un-arranging [of being as becoming the same management file name by NFS server 12-i which is different conversely in spite of being different file data substance **** /, and] may arise, and file access competition may arise between the NFS servers 12-i. / management file names differing by NFS server 12-i from which it differs here although such NFS server 12-i is the same file data substance when a file name is managed uniquely, respectively
[0025] The above-mentioned name server 13 solves such un-arranging. That is, in this name server 13, by carrying out unitary management of the meta-information access from NFS server 12-i, management file name space in all NFS server 12-i is set to one, and the file access competition between the NFS servers 12-i is avoided. Therefore, the reliability of file sharing by this system 1 will improve sharply by offering this name server 13.
[0026] in addition, it is shown in drawing 1 -- as -- this name server 13 -- heterologies (down), such as failure, -- offering -- present -- business and the object for reserves exist moreover, in hardware also about these name servers 13 Interface card 13c which equips an interface with the same composition as NFS server 12-i (refer to drawing 5), i.e., CPU13a, memory (primary storage) 13b, and the internal network 14 is mounted. The function as a name server 13 mentioned above is realized by CPU13a's reading the name server software (program) memorized by memory 13b also in this case, and operating.
[0027] Next, the redirector 11 of the above [the above-mentioned shared memory 15], NFS server 12-i, and a name server 13 are accessible memory through the internal network 14, respectively. for example, certain NFS server 12-i -- or -- present, when the name server 13 of business is downed (at the time of obstacle generating) the NFS server 12-i -- or -- present -- the information for taking over processing of the name server 13 of business to the name server 13 other NFS server 12-k (it being k!=i at k= 1 - n), or for reserves -- (Following and taking over information) etc. -- server 12-i and 13 -- it is held independently at memory card (Shared Memory Card) 15-1 - 15-m (m is the natural number) (refer to drawing 6 and drawing 8) (backup)
[0028] That is, each above-mentioned NFS server 12-i (CPU12a) or the name server 13 (CPU13a) will have the function as the taking over information Records Department 121 (131) (refer to drawing 5) which succeeds information required to offer at the time of a heterology and take over processing to the name server 13 other NFS server 12-i or for reserves, and is recorded on a shared memory 15 as information.
[0029] in addition, the down of NFS server 12-i is typically shown in drawing 6 -- as -- present -- the name server 13 (CPU13a) of business performs the NFS server surveillance demon 132 in the background -- supervising -- present -- as typically shown in drawing 8 , the down of the name server 13 of business is supervised because the name server 13 (CPU13a) for reserves performs the name server surveillance demon 133 in the background
[0030] And as typically shown in drawing 7 , when the down of NFS server 12-i is detected (Step S1) As opposed to NFS server 12-k (for example, NFS server 12-k with a light load) other than NFS server 12-i to which the name server 13 (CPU13a) of business was downed present -- While directing to succeed processing of downed NFS server 12-i, the down of NFS server 12-i is notified to a redirector 11 (Step S2).
[0031] NFS server 12-k (CPU12a) which received taking over directions accesses a shared memory 15 through the internal network 14 by this, and processing of NFS server 12-i which read the taking over information backed up there and was downed is succeeded (Step S3). On

the other hand, a redirector 11 (network processor 11c) is made not to transmit the request by the request transfer section 112 by receiving the above-mentioned notice from a name server 13 to downed NFS server 12-i at this time.
[0032] That is, the name server 13 (CPU13a) in this case The malfunction detection section 134 (refer to drawing 6) which detects the heterology of NFS server 12-i, If the heterology of NFS server 12-i is detected in this malfunction detection section 134 It will have the function as the taking over directions section 135 (refer to drawing 6) which gives [succeeding the processing of NFS server 12-i which carried out the heterology to other NFS server 12-k other than the NFS server 12-j based on the taking over information on a shared memory 15, and] taking over directions.
[0033] Thus, with this operation gestalt, since the name server 13 other NFS server 12-k and for reserves can succeed processing even if NFS server 12-i and a name server 13 are downed, file processing normal as a storage system 1 can be continued, and obstacle-proof nature improves sharply. In addition, although this example is explanation about redundant-izing of NFS server 12-i or a name server 13, of course, it is also possible to redundancy-ize a redirector 11 similarly. moreover -- present -- you may make it take over to either of the NFS server 12-i depending on the case about the taking over at the time of the name server 13 down of business

[0034] Next, the redirector 11 at the time of the request transfer to NFS server 12-i from the redirector 11 mentioned above and the concrete processing by NFS server 12-i are explained. A redirector 11 will analyze the file access request in the request analysis section 111, if the file access request for writing in a certain file data from a client 2 is received.
[0035] In addition, as shown in drawing 12 , the above-mentioned "file access request" has the header unit 21 which consists of physical-layer header (Phy Header) 21a, IP header (Internet Protocol Header) 21b, TCP header (Transmission Control Protocol Header) 21c, and NFS header 21d etc., and the real file data section 22 in which the substance (real file data) of file data which should actually be written in a secondary storage 17 was stored, and changes.
[0036] And the request analysis section 111 asks for the boundary of the position where the real file data under above-mentioned file access request starts, i.e., a header unit 21 and the real file data section 22, as header offset value [number-of-bits ("a") etc. 23 from boundary information;, for example, a head,], as typically shown in drawing 9 . The boundary information 23 searched for is notified to the request transfer section 112, and the request transfer section 112 appends the boundary information 23 notified to the above-mentioned file access request (addition), and sends it to NFS server 12-i of the destination.
[0037] That is, as shown in drawing 2 , the above-mentioned request analysis section 111 The function as header offset value analysis section 111a to calculate the header offset value 23 which analyzes the received file access request and expresses the boundary position of the header unit 21 of the file access request, and the real file data section 22, It will have the function as header offset pricing Kabe 111b which adds the header offset value 23 acquired by this header offset value analysis section 111a to the file access request to which it is transmitted to NFS server 12-i.
[0038] Then, in the NFS server 12-i side, it is based on the header offset value 23 added by the redirector 11 side like ****. The NIC (Network Interface Card) driver (network driver) 122 The real file data section 22 and the other field The start address of (a header unit 21) can be assigned to the page boundary [the page boundary (another field); buffer (mbuf) 123,124] of the message treated within the kernel of a high order layer (NFS processing layer) (kernel high order layer), respectively (refer to drawing 10).
[0039] When are done in this way and a file access request reaches the file system section 125 of a kernel high order layer as typically shown, for example in drawing 11 , it becomes possible to move data to the file system buffer 126 only by changing the start address (pointer) of the real file data section 22 for the pointer to the file system buffer 126, without generating a copy of data (map change) (the zero copy in a kernel is realized). Therefore, DMA (Direct Memory Access) can also be performed at high speed, and can improve sharply the processing speed and the throughput of NFS server 12-i.

[0040] In addition, although making it ask by the NIC driver 122 side is also thought of, the boundary of a header unit 21 and the real file data section 22 In this case, since the amount of processings (header analysis) of the NIC driver 122 increases (the NIC driver 122 usually performs only analysis of physical-layer header 21a), The direction which asked for the boundary by the redirector 11 side which has the analysis feature (request analysis section 111) of a header unit 21 from the first as mentioned above The kernel zero copy in a high order layer (NFS processing layer) can be realized, without increasing the throughput in a NIC driver layer (** which does not cause a throughput fall).
[0041] According to the storage system 1 of this operation gestalt, as mentioned above, to the system 1 interior By having formed a redirector 11, two or more NFS server 12-i, the name server 13, the shared memory 15, and the secondary storage 17, and having made these the composition connected in the high-speed internal network 14 Since NFS server 12-i and a secondary storage 17 can be extended easily if needed and it moreover is not necessary to maintain NFS server 12-i separately (maintenance) The performance (for example, it can respond to 10GbpsLAN) and capacity scalability which can fully respond by the low cost to band expansion of the external network 3 are securable.
[0042] Especially with the operation gestalt mentioned above, a redirector 11 so that processing may be uniformly assigned to each NFS server 12-i according to the load of each NFS server 12-i control or Assign the processing to the request about the same file to the same NFS12-i, or Since it answers by the reply message which carried out the cache by the redirector [not NFS server 12-i but] 11 side to the request about the high file of access frequency The processing speed and performance are improving by leaps and bounds, and the performance and the capacity scalability which can certainly respond are realized to 10GbpsLAN.
[0043] (B) NFS server 12-i with a larger capacity of memory 12b than other NFS server 12-i is arranged as cache server 12' (refer to drawing 1), and you may make it the file access in this cache server 12' return a response only by the R/W to memory 12b fundamentally in the system 1 in which the 1st modification carried out explanation **** at a client 2 side.
[0044] And about the high file of the access frequency whose request frequency within a fixed period is more than the number of times of predetermined (threshold), the cache is carried out to memory 12b of cache server 12', and it is made for cache server 12' to answer. Specifically, first, the access frequency of each file is supervised by the redirector 11 side, and about access to a file with access frequency higher than a certain threshold, directions are given to a name server 13, NFS server 12-i, and cache server 12' from redirector 12-i so that it may process by cache server 12-i.
[0045] That is, at this time, a redirector 11 (request transfer section 112) will transmit the request about the high file of access frequency to cache server 12'. Thereby, since access to the high file of access frequency is processed within cache server 12', without accessing a secondary storage 17, it contributes to the large improvement in the processing speed as a storage system 1, and a throughput greatly.
[0046] on the other hand, if the access frequency to the high file of the above-mentioned access frequency falls, a redirector 11 directs to assign suitable (for example, a load -- light) NFS server 12-i, and to shift processing to a name server 13, NFS server 12-i, and cache server 12' (if the access frequency to the file by which the cache is carried out to memory 12b of cache server 12' becomes below the number of times of predetermined)
[0047] That is, a redirector 11 (request transfer section 112) changes the destination of a request into NFS server 12-i other than cache server 12' in this case. While it is avoided that the cache of the file from which access frequency has fallen continues being forever carried out to cache server 12' by this, consequently it can cut down memory space required for cache server 12', it can give a margin to processing by cache server 12', and can improve the throughput.
[0048] (C) It is also possible to enable it to access a secondary storage 17 also from the external node 19 by [which is explanation of the 2nd modification] connecting the above-mentioned secondary storage 17 with a name server 13 and NFS server 12-i by FC switch 18 course (a secondary-storage network being built), and connecting the FC switch 18 and the

external node 19, as shown in drawing 13 . However, the file system operated by the external node 19 in this case needs to be the same as the file system in the storage system 1.
[0049] Although a certain mediation control is needed about access from the external node 19 by doing in this way in order to avoid the access competition with NFS server 12-i of the system 1 interior, access to the file in this storage system 1 from the external node 19 is attained. Since the management file name in a system 1 will be followed also about access from the external node 19 if access to a name server 13 from the external node 19 is permitted as it corrects, for example, is shown in drawing 14 , the file access from the external node 19 can be performed without needing the above-mentioned mediation control. In addition, although it is the case where access to the name server 13 in internal network 14 course is permitted, in this drawing 14 , you may make it, permit access by secondary-storage network (FC switch 18) course in drawing 13 , of course.
[0050] Moreover, as shown, for example in drawing 15 , you may communalize NFS server 12-i and the external node 19. That is, it constitutes so that file processing according to the request directly received from the external network 3 in NFS server 12-i may be performed to a secondary storage 17. When NFS server 12-i functions as a file server of the storage system 1 mentioned above by this when a certain client 2 has accessed by the redirector 11 course, and NFS server 12-i has been accessed directly, it will function as a usual file server which answers without going via a redirector 11. That is, the both sides of access by the NFS server 12-i course from a client 2 and the direct access which does not go via NFS server 12-i are permitted.
[0051] The direct access from the outside of any case becomes possible, and fusion to other storage architecture [SAN (Storage Area Network) etc.] can be realized. In addition, in drawing 14 and drawing 15 , a sign 20 expresses the network disk adapter which equips the interface of the internal network 14 and a secondary storage 17.
[0052] Moreover, although the shared memory 15 mentioned above is omitted with the composition shown in drawing 13 - drawing 15 , of course, it may be equipped. If it does in this way, also in the composition shown in drawing 13 - drawing 15 , the same backup processing as the above will be attained.
(D) In addition, although the operation gestalt mentioned above, in addition, explained the case where Gigabit Ethernet was applied as Infiniband and an external network 3 as an internal network 14, of course, it is also possible to carry out a system construction using other high-speed networks other than these.
[0053] Moreover, it is not necessary to necessarily offer an above-mentioned name server 13 and an above-mentioned shared memory 15, and even if it omits these either or both sides, the purpose of this invention is attained enough. Furthermore, although NFS is applied to the file server with the operation gestalt mentioned above, this invention is not limited to this but, of course, it is also possible to apply other file systems.
[0054] Moreover, although premised on the capacity (transmission speed) of the internal network 14 being about 4-10Gbps with the operation gestalt mentioned above, this speed can respond, if it changes suitably according to band expansion of the external network 3. And this invention is not limited to the operation gestalt mentioned above, but in the range which does not deviate from the meaning of this invention besides the above, can deform variously and can be carried out.
[0055] (E) Additional remark [additional remark 1] Storage which can memorize file data, Two or more file servers which perform file processing according to the request to this storage, The transfer processing to this file server of the request received from a client through an external network, The file server management node which carries out unitary management of the response processing to this client to this request, The storage system characterized by having offered the internal network which interconnects this storage, this file server, and this file server management node possible [communication], and being constituted.
[0056] [Additional remark 2] Storage system of the additional remark 1 publication characterized by connecting the name server which carries out unitary management of the file data name which this file server treats to this internal network.

[Additional remark 3] Storage system of the additional remark 1 publication characterized by connecting the shared memory with accessible this file server management node and this file server to this internal network.
[0057] [Additional remark 4] Storage system of the additional remark 2 publication characterized by connecting the shared memory with accessible this file server management node, this file server, and this name server to this internal network.
[Additional remark 5] Storage system given in any 1 term of additional remarks 1-4 characterized by having offered the request analysis section in which this file server management node analyzes the content of this request, and the request transfer section which transmits this request to a specific file server according to the analysis result of this request analysis section.
[0058] [Additional remark 6] Storage system of the additional remark 5 publication which offers the transfer history Records Department where this file server management node records the transfer history of the request of the past by this request transfer section, and is characterized by being constituted so that this request transfer section may transmit the request to the file data of the same file data name to the same file server based on this transfer history of this transfer history Records Department.
[0059] [Additional remark 7] Storage system of the additional remark 5 publication which this file server management node offers the load Monitoring Department which supervises the load of this file server, and is characterized by constituting this request transfer section so that this request may be transmitted to the low file server of a load based on the surveillance result in this load Monitoring Department.
[0060] [Additional remark 8] Storage system of the additional remark 5 publication which offers the primary storage to which at least one in this file server can carry out the cache of the file data in this storage, and is characterized by being constituted as a cache server which performs file processing according to this request in this primary storage.
[Additional remark 9] Storage system of the additional remark 8 publication characterized by being constituted so that this request transfer section may transmit the request about file data with the above-mentioned high request frequency to this cache server while this primary storage of this cache server was constituted so that the request frequency within a fixed period might carry out the cache of the file data more than the number of times of predetermined.
[0061] [Additional remark 10] Storage system of the additional remark 9 publication characterized by being constituted so that the destination of this request may be changed into file servers other than this cache server when the request frequency to this file data by which the cache of this request transfer section is carried out to this primary storage of this cache server became below the number of times of predetermined.
[0062] [Additional remark 11] The storage system of the additional remark 5 publication which carries out [having offered the header offset value analysis section which calculates the header offset value as which this request analysis section analyzes this request, and expresses the boundary position of the header unit of the request concerned, and the real file data section, and header offset pricing Kabe who add this header offset value acquired in this header offset value analysis section to this request to which it is transmitted to this file server, and] as the feature.

[0063] [Additional remark 12] Storage system of the additional remark 11 publication characterized by having offered the network driver by which this file server copies this header unit and these data division of this request to the field to which the messages treated in a kernel high order layer, respectively differ based on this header offset value added to this request.
[Additional remark 13] Storage system given in any 1 term of additional remarks 1-12 characterized by having offered the cache section to which this file server management node carries out the cache of the response message to this client about specific file data with high request frequency, and the response section which returns the applicable response message of this cache section to this client as this request is a thing about this specific file data.
[0064] [Additional remark 14] Storage system given in the additional remark 3 or additional remark 4 characterized by having offered the taking over information Records Department which

succeeds information required for this file server to offer at the time of a heterology, and take over processing to other file servers, and records on this shared memory as information.
[Additional remark 15] Storage system of the additional remark 14 publication characterized by preparing the malfunction detection section which detects the heterology of this file server, and the taking over directions section which will give [succeeding processing of this unusual file server based on this taking over information on this shared memory to other file servers other than the file server (henceforth an unusual file server) concerned and] taking over directions if the heterology of this file server is detected in this malfunction detection section.
[0065] [Additional remark 16] Storage system given in any 1 term of additional remarks 1-15 characterized by constituting this storage so that access from an external node may be permitted.
[Additional remark 17] Storage system given in any 1 term of additional remarks 2-15 characterized by constituting this name server so that access from an external node may be permitted.
[0066] [Additional remark 18] Storage system given in any 1 term of additional remarks 1-17 characterized by being constituted so that this file server may perform file processing according to the request received directly from this external network to this storage.
[0067]
[Effect of the Invention] Two or more file servers which perform file processing according to the request to storage according to the storage system of this invention as explained in full detail above, Since the internal network which interconnects the file server management node which carries out unitary management of the processing of each file server, and storage, a file server and a file server management node possible [communication] is offered Since a file server and storage can be extended easily if needed and it moreover is not necessary to maintain each file server separately (maintenance) The performance and capacity scalability which can fully respond by the low cost to band expansion of an external network are securable.
[0068] And if the name server which carries out unitary management of the file data name which each above-mentioned file server treats to the above-mentioned internal network is connected, since it can prevent that file access competition arises between file servers, it contributes to the improvement in reliability of file sharing greatly. Moreover, though an obstacle occurs in some file servers, since processing normal as a storage system is continuable by memorizing the taking over information on the file server which connected the shared memory to the above-mentioned internal network, and was offered on it at this shared memory at the time of obstacle generating at any time, obstacle-proof nature improves sharply.
[0069] Furthermore, if a file server management node may be constituted so that the request to the file data of the same file data name may be transmitted to the same file server based on the transfer history of the past request, and it does in this way, its processing speed will improve sharply. Moreover, if the load of a file server is supervised, you may make it transmit a request to the low file server of a load and it does in this way, since a request can be uniformly transmitted to each file server, what a load concentrates on some file servers and causes an obstacle can be prevented certainly, and reliability improves it sharply.
[0070] Furthermore, about file data with high request frequency, if the cache is carried out to the primary storage of a cache server and it is made to process by the cache server, since the access frequency to storage is sharply reducible, processing speed and processability ability improve further. And if the request frequency to the file data by which the cache is carried out to the primary storage of a cache server in this case becomes below the number of times of predetermined and it will be made to process by file servers other than a cache server Since it is avoided that the low file data of request frequency continues being held forever at a cache server While memory space required for the primary storage of a cache server is reducible, a margin can be given to processing by the cache server and the throughput can be improved.
[0071] Moreover, in a file server management node, the header offset value showing the boundary position of the header unit of a request and the real file data section is calculated, and if the header offset value is added to a request and it is made to transmit to a file server, in the network driver of a file server, the header unit and data division of a request can be copied to

the field to which the messages treated in a kernel high order layer, respectively differ based on the header offset value. Therefore, the zero copy in a kernel can be realized and the processing speed and the throughput of a file server can be improved sharply.
[0072] Furthermore, if the response message which carries out the cache of the response message to the client about specific file data with high request frequency, and carried out the cache to the received request being a thing about the file data in the above-mentioned file server management node is returned to a client, since there will be no need of transmitting a request to a file server, the speed of response to a client will improve sharply, and the processing speed and the throughput as this system will improve by leaps and bounds.
[0073] Moreover, if the above-mentioned storage may be constituted so that access from an external node may be permitted, and it does in this way, fusion to other storage architecture will be attained. If access to the above-mentioned name server [node / external] is permitted here, the file access from an external node will become possible, without needing file access mediation control with the above-mentioned file server and an external node.
[0074] Furthermore, if the above-mentioned file server may be constituted so that file processing according to the request received directly may be performed from the above-mentioned external network to storage, and it does in this way, since the both sides of access via the file server from a client and the direct access which does not go via a file server are permissible, fusion to other storage architecture is attained also in this case.

[Translation done.]

* NOTICES *

Japan Patent Office is not responsible for any
damages caused by the use of this translation.

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1] It is the block diagram showing the storage structure of a system (storage architecture) as 1 operation gestalt of this invention.
[Drawing 2] It is the block diagram showing the composition of a redirector shown in drawing 1 .
[Drawing 3] It is a block diagram for explaining the case where the cache of the meta-information is carried out by the redirector shown in drawing 1 .
[Drawing 4] It is a block diagram for explaining the case where a reply message is returned by the redirector shown in drawing 1 .
[Drawing 5] It is the block diagram showing the composition of an NFS server (name server) shown in drawing 1 .
[Drawing 6] It is a block diagram for explaining the case where the taking over information on an NFS server is backed up to the shared memory shown in drawing 1 .
[Drawing 7] It is a block diagram for explaining the case where processing of the NFS server which was backed up by the shared memory shown in drawing 6 and which succeeded and was downed based on information is succeeded.
[Drawing 8] It is a block diagram for explaining the case where the taking over information on a name server is backed up to the shared memory shown in drawing 1 .
[Drawing 9] It is a block diagram for explaining the case where the boundary information on a file access request is searched for in the redirector shown in drawing 1 .
[Drawing 10] It is a block diagram for explaining the case where the zero copy in a kernel is realized based on the boundary information on the file access request shown in drawing 9 in the NFS server shown in drawing 1 .
[Drawing 11] It is a block diagram for explaining the case where the zero copy in a kernel is realized based on the boundary information on the file access request shown in drawing 9 in the NFS server shown in drawing 1 .
[Drawing 12] It is drawing showing the example of a format of the file access request shown in drawing 9 - drawing 11 .
[Drawing 13] It is the block diagram showing the composition in the case of permitting access to the secondary storage from an external node in the storage system shown in drawing 1 .
[Drawing 14] It is the block diagram showing the composition in the case of permitting access to the name server from an external node in the storage system shown in drawing 1 .
[Drawing 15] It is a composition **** block diagram in the case of permitting access via the redirector from an external node, and direct access in the storage system shown in drawing 1 .
[Drawing 16] It is a block diagram for explaining the conventional technique of realizing file sharing between two or more nodes (client) on a network.
[Description of Notations]
1 Storage System
2 Client
3 External Network (Gigabit Ethernet)
11 Redirector (Redirector; File Server Management Node)
11a Gigabit Ethernet card

11b In FINI band card
11c Network processor
11d Memory (primary storage)
11e PCI (Peripheral Component Interconnect) bus
11f Cache memory (cache section)
12-1 - 12-n NFS (Network File System) server (file server)
12' Cache server
12a,13a CPU(Central Processing Unit)
12b, 13b Memory (primary storage)
12c, 13c Interface card (IB-IF)
13 Name Server
14 High-speed (Interior) Network [in FINI Band (Infiniband)] Switch
15 Shared Memory (Shared Memory)
15-1 - 15-m Memory card (Shared Memory Card)
16 IB-FC Card
17 Secondary Storage
18 FC Switch
19 External Node
20 Network Disk Adapter
21 Header Unit
21a Physical-layer header (Phy Header)
21b IP header (Internet Protocol Header)
21c TCP header (Transmission Control Protocol Header)
21d NFS header
22 Real File Data Section
23 Header Offset Value (Boundary Information)
111 Request Analysis Section
111a Header offset value analysis section
111b Header offset pricing Kabe
112 Request Transfer Section
113 Transfer History Records Department
114 Load Monitoring Department
115 NFS Server Load Surveillance Demon (Daemon)
116 Response Section
121,131 Taking over information Records Department
122 NIC (Network Interface Card) Driver (Network Driver)
123,124 Buffer (mbuf)
125 File System Section
126 File System Buffer
132 NFS Server Surveillance Demon
133 Name Server Surveillance Demon
134 Malfunction Detection Section
135 Taking over Directions Section

[Translation done.]